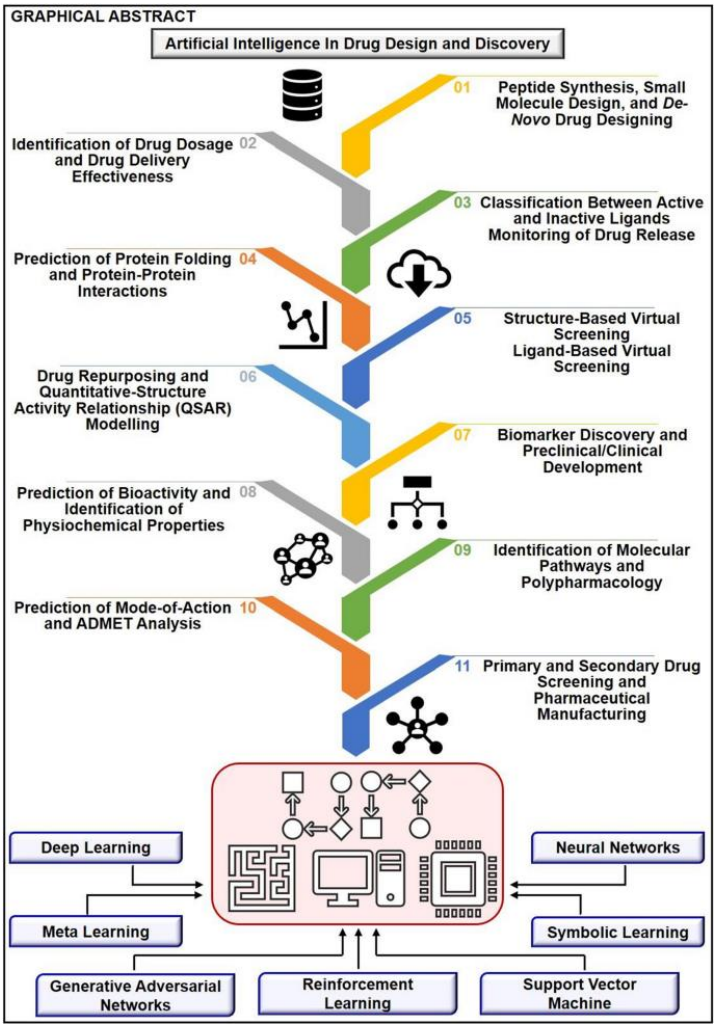
1.人工智能药物发现(AIDD)简介

人工智能药物发现（AIDD）是一种将人工智能（AI）技术应用于药物研发的方法。AIDD 利用 AI 算法来分析大规模的分子结构数据，以帮助预测分子间的相互作用及其对于疾病的治疗效果。



AIDD 涉及多种 AI 技术，如机器学习、深度学习、神经网络和自然语言处理等。其中，机器学习是 AIDD 的核心技术之一，它可以从大量的数据中学习模式和规律，并根据这些规律预测新的结果。通过对已知药物的化学结构和生物活性进行机器学习，可以为新药物的设计和开发提供重要的指导。

AIDD 还可以用于虚拟筛选，即通过计算机模拟来预测分子与目标蛋白之间的相互作用，以便确定潜在的治疗靶点和药物化合物。这种虚拟筛选方法可以大大减少实验室筛选的时间和成本，从而加快新药物的发现和研发进程。

此外，AIDD 也能用于药物副作用的预测和控制。通过分析药物和生物体之间的相互作用，AIDD 可以预测药物的副作用和毒性，避免不必要的实验和动物试验，帮助研究人员设计更加安全的药物。

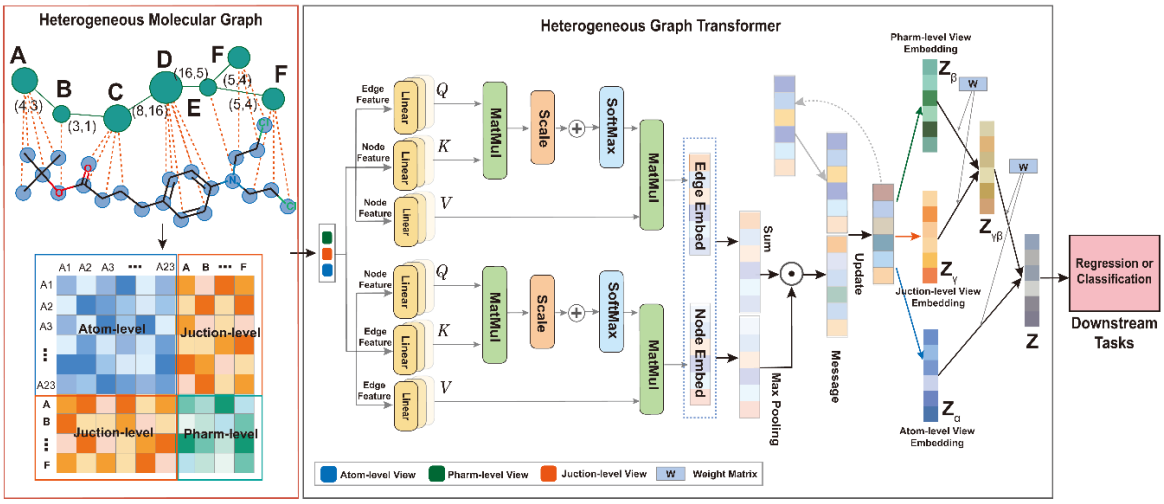
虽然 AIDD 在药物研发领域已经取得了一些显著的成果，但它还面临着一些技术和法律上的挑战。例如，由于药物研发是一项涉及人类生命健康的重要领域，因此在利用 AIDD 时需要遵循相关的法律和道德规范，确保药物的安全性和有效性。此外，尽管 AIDD 可以帮助研究人员加速药物研发进程，但它并不能完全替代实验室实验和动物试验，因此需要综合考虑各种因素来做出决策。

总之，AIDD 是一项具有广阔前景的技术，可以加速药物研发进程，为治疗疾病提供更多的选择和机会。随着技术的不断发展和完善，相信 AIDD 将在未来发挥更加重要的作用。

2.机器学习和深度学习在药物发现领域的应用

2.1 分子属性预测与优化

分子属性预测与优化是生物信息学和化学领域中的重要概念，它涉及预测分子的性质和结构，然后对这些分子进行优化以满足特定的需求。分子属性预测，是指使用计算方法、实验数据或机器学习技术来预测分子的性质和特性。例如，可以预测分子的物理和化学属性，如溶解度、毒性、药效、反应性等，以便更好地了解它们在不同环境中的行为。分子优化，是指通过改变分子的结构或性质，以满足特定的目标或需求。例如，在药物设计中，分子优化可以用于改进候选药物的活性、选择性或药代动力学性质。这通常包括分子的结构修改或参数调整，以获得更理想的特性。



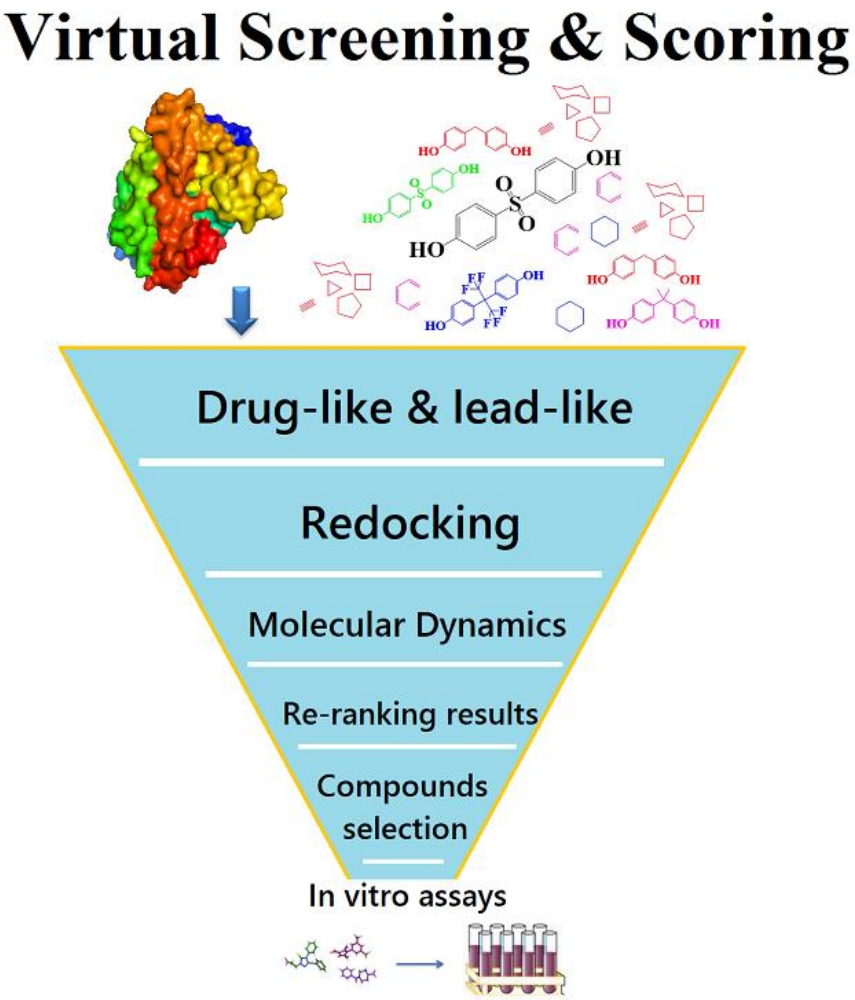
传统的分子属性预测方法通常基于物理化学模型，但这些方法通常复杂且计算成本高昂。这为机器学习和深度学习带来了机会，因为它们可以从大量的分子数据中学习模式，从而更有效地预测分子属性。在分子属性预测方面，机器学习和深度学习模型可以预测分子的物理化学性质，如溶解度、毒性、生物活性等。这对于药物设计和毒性评估非常重要。此外，分子优化也受益于机器学

习。通过优化分子的结构，可以改进其性质，例如提高药物分子的活性或改进材料的性能。机器学习模型可以协助进行结构优化，从而减少实验的试错次数。

机器学习和深度学习已经成为生物信息学和化学领域不可或缺的工具。它们为分子属性预测与优化带来了高效性和准确性，加速了新药物、新材料和生物分子的研究和开发。未来，我们可以期待更多创新和应用，将这些技术推向新的高度，以解决重要的医疗、环境和材料挑战。

2.2 虚拟筛选

虚拟筛选（Virtual Screening）是一种计算化学和生物信息学的方法，用于在大型分子数据库中筛选潜在的活性化合物或分子，以寻找可能与特定生物分子（如蛋白质）相互作用并表现出所需活性的候选物质。虚拟筛选是生物信息科学和药物研究领域的关键组成部分，用于识别潜在的候选化合物或生物分子，以加速药物发现和药物研发过程。近年来，机器学习和深度学习技术已经在虚拟筛选中发挥了重要作用，为科研人员提供了更强大的工具来预测化合物的性质、生物活性和药效，以及加速新药物的发现。



通过构建机器学习模型，我们可以分析已知的生物活性数据，从而预测化合物的潜在活性。这有助于筛选出最有希望的化合物，减少实验室试验的成本和时间。其次，深度学习技术在虚拟筛选中具有显著的优势。深度神经网络（DNN）可以处理大规模和高维度的分子数据，包括分子结构、蛋白质序列和化学特性。这种模型的复杂性允许它们更好地捕捉分子之间的相互作用和关联，从而提供更准确的预测。例如，卷积神经网络（CNN）可以用于图像识别，而图卷积神经网络（GCN）可用于分子图数据的分析。这些技术已经成功用于药物发现和蛋白质-蛋白质相互作用的预测。

然而，在利用机器学习和深度学习进行虚拟筛选时，也存在一些挑战。首先，数据质量和数量对模型的性能至关重要。需要大量准确的数据来训练和验证模型，而不准确或有偏差的数据可能导致模型的预测不准确。此外，解释性是一个重要问题，深度学习模型通常被视为“黑盒”，难以解释其决策过程，这在药物研究中可能引发监管和伦理方面的问题。最后，虚拟筛选的结果需要在实验室中进行验证，以确认候选分子的生物活性和安全性。

2.3 药物副作用预测与安全评估

药物安全评估是药物研究与开发中至关重要的环节，因为它有助于识别潜在的不良副作用并提前预测药物的安全性。在这个领域，机器学习和深度学习技术已经成为强大的工具，能够加速药物副作用的预测和药物安全性评估。

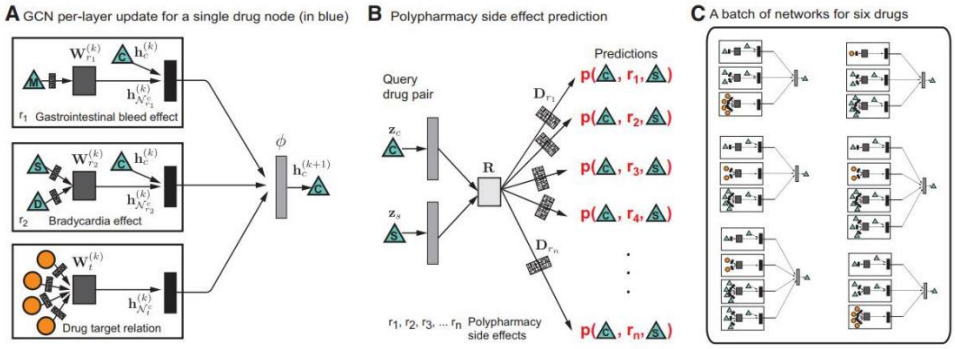


Fig. 3. Overview of *Decagon* model architecture. (**A**) An *Decagon* encoder. Shown is a per-layer update for a single graph node (a drug node representing Ciprofloxacin based on the small example input graph in Fig. 1). Hidden state activations from neighboring nodes $\mathcal{N}_r^c$ are gathered and then transformed for each relation type r individually (i.e. gastrointestinal bleed, bradycardia and drug target relation). The resulting representation is accumulated in a (normalized) sum and passed through a non-linear activation function ϕ (i.e. ReLU) to produce hidden state of node v_c in the $(k+1)$-th layer, $\mathbf{h}_c^{(k+1)}$. This per-node update is computed in parallel with shared parameters across the whole graph. (**B**) For every relation, *Decagon* decoder takes pairs of embeddings (e.g. hidden node representations $\mathbf{z}_c$ and $\mathbf{z}_s$ representing Ciprofloxacin and Simvastatin) and produces a score for every (potential) edge in the graph. Shown is the decoder for poypharmacy side effects relation types. (**C**) A batch of neural networks that compute embeddings of six drug nodes in the input graph. In *Decagon*, neural networks differ from node to node but they all share the same set of relation-specific trainable parameters [i.e. the parameters of the encoder and decoder; see Equations (1) and (2)]. That is, rectangles with the same shading patterns share parameters, and thin rectangles with black and white shading pattern denote densely connected neural layers

计算机方法可以分析大规模的临床数据、生物信息学数据以及文献信息，从而发现与特定药物相关的潜在副作用。例如，支持向量机（SVM）、随机森林（RF）和深度神经网络（DNN）等算法可以用于构建预测模型，根据分子结构、药物代谢途径和生物活性等因素来评估药物的毒副作用。

深度学习技术在这一领域中表现出色。深度神经网络可以处理高维度和复杂的生物数据，包括基因表达、蛋白质-蛋白质相互作用和药物-蛋白质相互作用数据。这使得它们能够更好地捕捉不同因素之间的复杂关系，提供更准确的副作

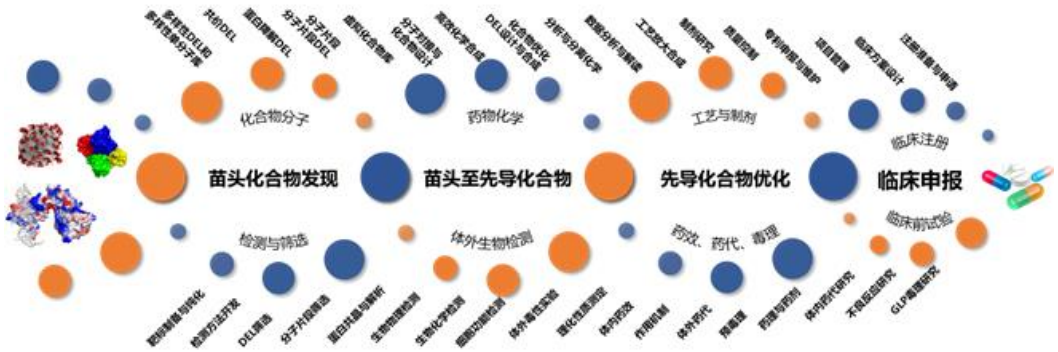
用预测。例如，递归神经网络（RNN）和卷积神经网络（CNN）可以用于处理序列数据和图数据，这些数据常常与药物副作用相关。

我们还可以利用这些算法用于药物安全性评估。这包括了对药物与不同患者个体之间的差异、药物相互作用以及代谢途径的研究。深度学习模型可以分析个体的遗传信息，以预测他们对某种药物的反应，从而为个体化治疗提供指导。

随着这些技术的不断改进和深入研究，我们可以期待更多的创新，以加速药物研发并提高患者的安全性。然而，需要解决数据质量、解释性和实验验证等方面的问题，以充分发挥机器学习和深度学习在药物安全性评估中的潜力。这将是生物信息科学研究的关键挑战和机遇之一。

2.4 新药分子设计

新药分子设计是药物研发领域的关键环节，旨在寻找具有特定治疗效果且安全的候选化合物。传统的药物设计方法通常是昂贵和耗时的，但机器学习和深度学习技术已经在这一领域取得了突破，为研究人员提供了更快速、高效和精确的工具。



计算机算法可以用于生成具有特定性质的分子结构，例如药物活性、药代动力学参数和毒性。生成对抗网络（GANs）是一种深度学习技术，可用于生成具有特定性质的化合物结构，允许研究人员有针对性地设计候选化合物。这加速了新药分子的开发，减少了试验和合成的时间和资源成本。

此外，对于新生成的药物分子，我们还需要对其性质、药效、不良反应等进行评估和预测。例如，通过深度学习模型进行药物-靶标互作预测，分析蛋白质序列、蛋白质-蛋白质相互作用和生物途径数据，以识别可能的治疗靶点，分析和确定分子的药效程度。预测药物的效应和副作用，这有助于筛选潜在的药物分子，以确保其具有足够的特异性和最小的不良影响，确定候选药物的有效性和安全性。

3. 工具介绍与安装

3.1 Anaconda3 / Pycharm / Jupyter Notebook 安装

（1）介绍

Python 是一种面向对象的解释型计算机程序设计语言，其使用，具有跨平台的特点，可以在 Linux、macOS 以及 Windows 系统中搭建环境并使用，其编写的代码在不同平台上运行时，几乎不需要做较大的改动，使用者无不受益于它的便捷性。此外，Python 的强大之处在于它的应用领域范围之广，遍及人工智能、科学计算、Web 开发、系统运维、大数据及云计算、金融、游戏开发等。实现其强大功能的前提，就是 Python 具有数量庞大且功能相对完善的标准库和第三方库。通过对库的引用，能够实现对不同领域业务的开发。然而，正是由于库的数量庞大，对于管理这些库以及对库作及时的维护成为既重要但复杂度又高的事情。

Anaconda（官方网站）就是可以便捷获取包且对包能够进行管理，同时对环境可以统一管理的发行版本。Anaconda 包含了 conda、Python 在内的超过 180 个科学包及其依赖项。Anaconda 具有如下特点：开源、安装过程简单、高性能使用 Python 和 R 语言、免费的社区支持、其特点的实现主要基于 Anaconda 拥有的 conda 包、环境管理器、1,000+开源库。如果日常工作或学习并不必要使用 1,000 多个库，那么可以考虑安装 Miniconda（下载界面请戳），这里不过多介绍 Miniconda 的安装及使用。

（2）Anaconda3 的下载与安装（Python3.8）

下载网址：https://repo.anaconda.com/archive/

下载好 annaconda 后，再对其进行安装。

安装可参考此博客：
https://blog.csdn.net/qq_44955003/article/details/120385651

（3）Pycharm 下载与安装

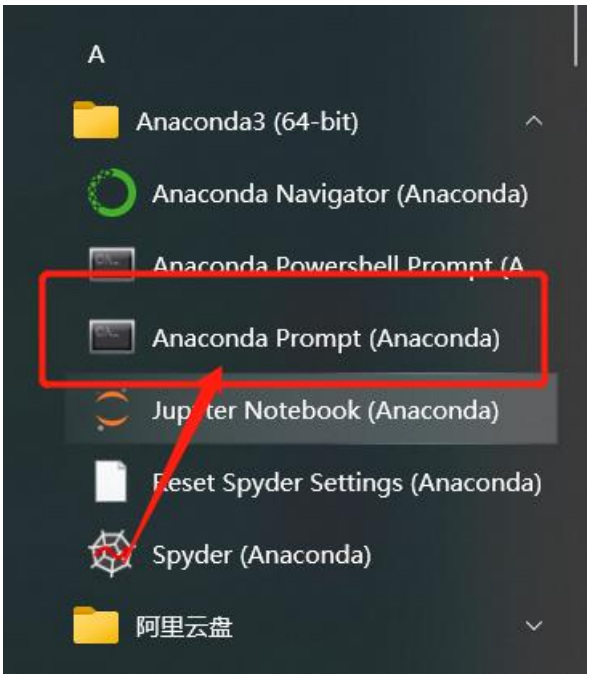
接下来安装 Pycharm，官网下载地址为：
https://www.jetbrains.com/pycharm/download/#section=windows

安装可参考此博客：
https://blog.csdn.net/qq_44955003/article/details/120385651

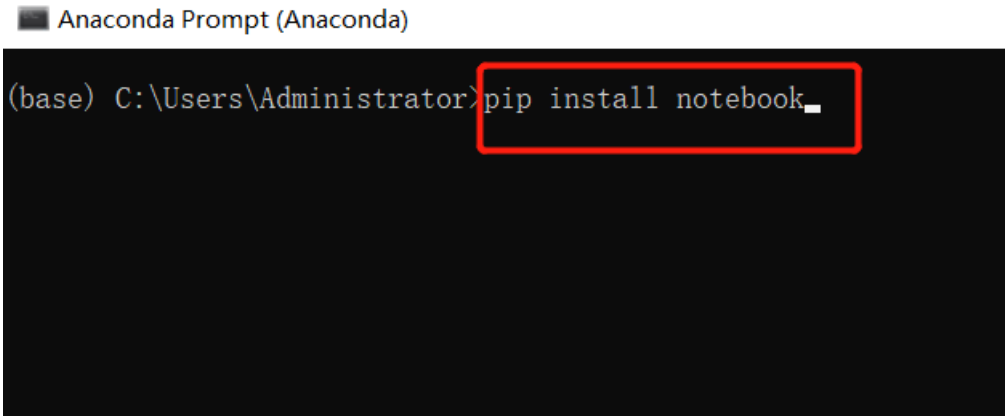
（4）Jupyter Notebook 下载与安装

Jupyter Notebook 是基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。为了更加方便的展示代码运行结果，因此我们需要在电脑上安装 Jupyter Notebook。

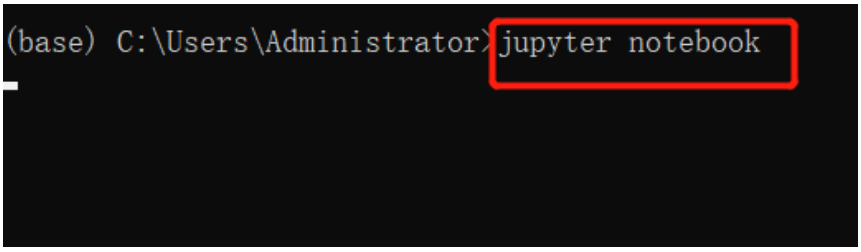
第一步：打开 Anaconda Prompt



第二步:在命令行输入 pip install notebook

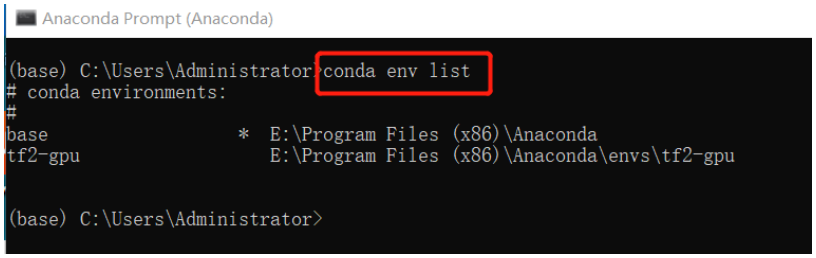


第三步：安装好之后，启动 Jupyter notebook

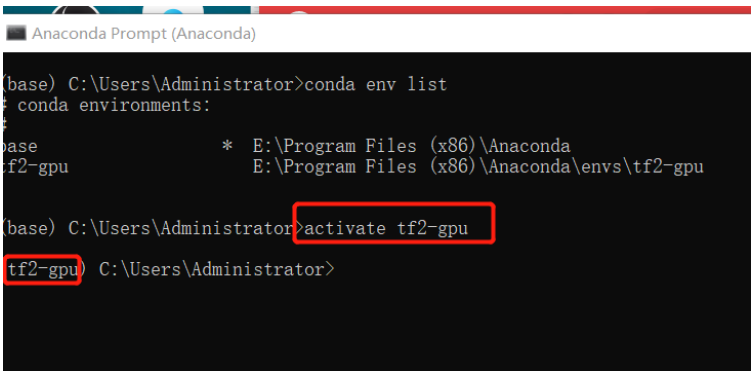


（5）如何在 Jupyter notebook 中添加新环境

同样的，打开 Anaconda Prompt。在命令行输入：conda env list



列出目前我们所创建的几个环境，通过 activate tf2-gpu 命令激活我们所要使用的环境。



这里我们发现，我们成功的从 base 环境进入到 tf2-gpu 的环境下。接下来，就是在改环境下安装内核，输入：conda install ipykernel

安装完成后，为 Jupyter notebook 添加环境，输入：

python -m ipykernel install --user --name tf2-gpu --display-name "前端展示的名称"

启动 Jupyter notebook，我们发现环境添加成功。这里演示的是在添加 TensorFlow 环境，pytorch 的环境添加也可按照这些步骤操作。

3.2 Python 基础

Python 是一门开源免费、通用型的脚本编程语言，它上手简单，功能强大，坚持「极简主义」。Python 类库（模块）极其丰富，这使得 Python 几乎无所不能，不管是传统的 Web 开发、PC 软件开发、Linux 运维，还是当下火热的机器学习、大数据分析、网络爬虫，Python 都能胜任。

3.3 Numpy 基础

NumPy（Numerical Python）是 Python 的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比 Python 自身的嵌套列表（nested list structure)结构要高效的多（该结构也可以用来表示矩阵（matrix)），支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。

3.4 Pandas 基础

pandas 是基于 NumPy 的一种工具，该工具是为解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas 提供了大量能使我们快速便捷地处理数据的函数和方法。

Pandas 是 Python 的核心数据分析支持库，提供了快速、灵活、明确的数据结构，旨在简单、直观地处理关系型、标记型数据。Pandas 的目标是成为 Python 数据分析实践与实战的必备高级工具，

Pandas 适用于处理以下类型的数据：

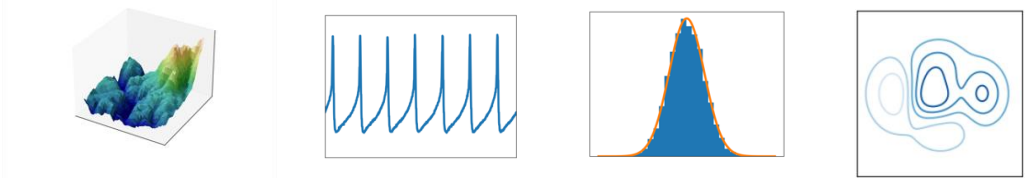
- 与 SQL 或 Excel 表类似的，含异构列的表格数据;
- 有序和无序（非固定频率）的时间序列数据;
- 带行列标签的矩阵数据，包括同构或异构型数据;
- 任意其它形式的观测、统计数据集，数据转入 Pandas 数据结构时不必事先标记。

数据结构

- Series：一维数组，与 Numpy 中的一维 array 类似。二者与 Python 基本的数据结构 List 也很相近。Series 如今能保存不同种数据类型，字符串、boolean 值、数字等都能保存在 Series 中。
- Time- Series：以时间为索引的 Series。
- DataFrame：二维的表格型数据结构。很多功能与 R 中的 data.frame 类似。可以将 DataFrame 理解为 Series 的容器。
- Panel ：三维的数组，可以理解为 DataFrame 的容器。

3.5 Matplotlib 基础

Matplotlib 是一个 Python 2D 绘图库，它以多种硬拷贝格式和跨平台的交互式环境生成出版物质量的图形。 Matplotlib 可用于 Python 脚本，Python 和 IPython Shell、Jupyter 笔记本，Web 应用程序服务器和四个图形用户界面工具包。



Matplotlib 尝试使容易的事情变得更容易，使困难的事情变得可能。 您只需几行代码就可以生成图表、直方图、功率谱、条形图、误差图、散点图等。 更多的示例，请参见基础绘图例子和示例陈列馆。

为了简单绘图，该 pyplot 模块提供了类似于 MATLAB 的界面，尤其是与 IPython 结合使用时。 对于高级用户，您可以通过面向对象的界面或 MATLAB 用户熟悉的一组功能来完全控制线型，字体属性，轴属性等。

3.6 scikit-learn 安装

Scikit-learn 是一个开源的机器学习库，它支持有监督和无监督的学习。它还提供了用于模型拟合，数据预处理，模型选择和评估以及许多其他实用程序的各种工具。

Scikit-learn 的安装：

操作系统：Windows

包管理器：conda

安装 conda(不需要管理员权限).

然后运行：

```
conda install scikit-learn
```

可以使用以下语句去检查

```
conda list scikit-learn # 查看 scikit-learn 安装的位置及安装的版本
conda list # 查看所有在虚拟环境中已下载的包
python -c "import sklearn; sklearn.show_versions()"
```

3.7 cuda / Pytorch 安装

3.7.1 Pytorch 介绍

PyTorch 是由 Facebook 人工智能研究小组开发的开源机器学习框架。它被广泛用于开发深度学习模型，并以其动态计算图而闻名，与 TensorFlow 等其他深度学习框架使用的静态计算图相比，它允许更大的灵活性和更快的原型化。

PyTorch 支持一系列任务，包括图像和语音识别、自然语言处理和计算机视觉。它构建在 Python 之上，为创建和训练神经网络提供了直观的界面。此外，PyTorch 拥有强大的社区和丰富的工具和库生态系统，使其成为研究和行业的热门选择。

PyTorch 的一些关键特性包括:

- 动态计算图:PyTorch 在执行操作时实时构建计算图，这使得模型设计更加灵活，原型制作速度更快。
- 自动微分:PyTorch 提供了自动微分，这使得在训练过程中计算梯度和优化模型变得很容易。
- GPU 加速:PyTorch 提供了对 GPU 加速的支持，这允许更快地训练深度学习模型。
- PyTorch 提供了一种将 Python 代码转换为序列化表示的方法，可以在不同的平台上有效地执行，包括移动和 web 平台。
- 分布式训练:PyTorch 支持跨多个 gpu 和机器的分布式训练，使其更容易将模型扩展到更大的数据集和复杂的架构。

总的来说，PyTorch 是一个强大而灵活的机器学习框架，近年来越来越受欢迎。它的动态计算图、自动微分和对 GPU 加速的支持使其非常适合深度学习的研发，而其工具和库的生态系统使其更容易入门和构建复杂的模型。

3.7.2 cuda / pytorch 的下载与安装

cuda 的安装可以参考这篇博客：
https://blog.csdn.net/m0_45447650/article/details/123704930

PyTorch 的官网：https://pytorch.org/。

PyTorch 的安装具体可以依照这篇博客：
https://blog.csdn.net/weixin_43828245/article/details/124779887

3.7.3 pytorch 的入门

见 pytorch.ipynb

3.8 RDKit 基础

3.8.1 RDKit 的介绍

RDKit 是一个用于化学信息学的开源工具包，基于对化合物 2D 和 3D 分子操作，利用机器学习方法进行化合物描述符生成，fingerprint 生成，化合物结构相似性计算，2D 和 3D 分子展示等。基于 Python 语言进行调取使用。

官网：http://www.rdkit.org

文档：http://rdkit.chenzhaoqiang.com/index.html

3.8.2 RDKit 下载与安装

这里介绍一种最快速的安装方法，由于 RDKit 是基于 python 语言使用的，所以可以在 anaconda 上快速进行 RDKit 的安装。

```
1.conda install -c rdkit rdkit
```

安装完成后，可以在 python 界面进行 import，来 check RDKit 是否安装成功。如果可能顺利 import，则说明安装成功。

```
2.import rdkit
```

3.8.3 RDKit 基础

具体见 RDKit.ipynb